Machine Learning

Stanford University
Professor Andrew Ng

Jordan Hong

May 23, 2020

Contents

1 Introduction

1.1 What is Machine Learning

1. Machine Learning
   - Grew out of work in Artificial Intelligence (AI)
   - New capabilities for computers
2. Examples:
   - database mining
   - applications can't programby hand (handwriting recognition, Natural Language Processing (NLP), Computer Vision)
   - Neuromorphic applications

3. Definition

   - Arthur Samuel(1959)

     > Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

   - Tom Mitchell(1998)

     > Well-posed Learning Problem: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

4. Machine Learning in this course:

   (a) Suupervised Learning

   (b) Unsupervised Learning

   (c) Others: reinforcement learning, recommender systems

   (d) Practical application techniques

1.2 Supervised Learning

In supervised learning, the *the right answer* is given. For example:

1. Regression: predict real-valued output.

2. Classification: predict discrete-valued output.

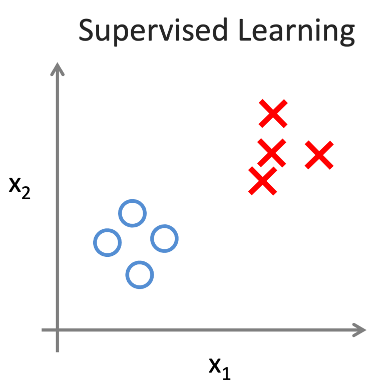

Figure 1: Supervised Learning

1.3 Unsupervised Learning

The right answer is not given, e.g. cocktail problem (distinguishing two voices from an audio file.)

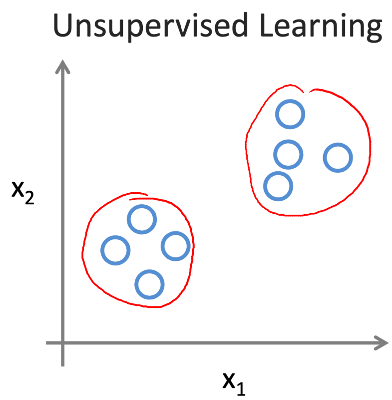

Figure 2: Unsupervised learning

2 Linear Regression with One Variable

2.1 Model Representation

2.1.1 Notations

For a training set:

- **m** = Number of training examples.
- **x** = “input” variable / features.
- **y** = “output” variables / “target” variable.
- **(x,y)** - one training example.
- **(x^i,y^i)** denotes the i^{th} training example

2.1.2 Hypothesis Function

A hypothesis function (h) maps input (x) to estimated output (y). How do we represent h?

$$\boxed{\textbf{Hypothesis Function} \quad h_\theta(x) = \theta_0 + \theta_1 x} \tag{1}$$

We can apply *Univariate linear regression* with respect to x.

2.2 Cost Function

Recall 1. The θ_is are parameters we have to choose. The intuition is is that we want to choose θ_i s such that h$_\theta$ is closest to y for our training examples (x,y).

$$\boxed{\textbf{Cost Function} \quad J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2} \tag{2}$$

Summary

1. **Hypothesis** $h_\theta(x) = \theta_0 + \theta_1 x$
2. **Parameters** θ_0, θ_1
3. **Cost Function** $J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$
4. **Goal** $\min_{\theta_0,\theta_1} J(\theta + 0, \theta_1)$

2.3 Gradient Descent

2.3.1 Intuition

1. We have some function $J(\theta_0, \theta_1)$, we want to $\min_{\theta_0,\theta_1} J(\theta_0, \theta_1)$
2. Outline: start with some θ_0, θ_1, keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we end up at a minimum.

2.3.2 Gradient Descent Algorithm

Algorithm
repeat until convergence{

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad \text{(for j=0 and j=1).}$$

}

Notes

1. the := denotes non-blocking assignment, i.e. simultaneously updates $\theta_0 and \theta_1$
2. We use the derivative to find a local minimum.
3. α denotes the learning rate. Gradient descent can converge to a local minimum even when the learning rate α is fixed. As we approach a local minimum, gradient descent will automatically take smaller steps. Therefore it is not needed to decrease α over time.

2.3.3 Gradient Descent with Linear Regression

Recall, we have:

1. Gradient Descent Algorithm:

   repeat until convergence{

   $$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad \text{(for j=0 and j=1).}$$

   }

2. Linear Regression Model:

   $$h_\theta(x) = \theta_0 + \theta_1 x$$

   $$\text{J}(\theta_0, \theta_1) = \tfrac{1}{2m} \textstyle\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

   .

We can substitute the above equations, which gives us:

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})$$

$$\theta_1 := \theta_1 - \alpha \tfrac{1}{m} \textstyle\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

.